Министерство образования Республики Беларусь
Белорусский государственный университет
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
ПРИКЛАДНЫХ ПРОБЛЕМ МАТЕМАТИКИ И ИНФОРМАТИКИ

УТВЕРЖДАЮ
Директор НИИ прикладных проблем
математики и информатики

Ю.С.Харин
«___» _____ 2022 г.

МЕТОДИКА
ОЦЕНКИ ЭНТРОПИИ ИСТОЧНИКОВ СЛУЧАЙНОСТИ

МИ.10127.10.02

Листов 31

Минск 2022

Предисловие

Настоящая методика испытаний предназначена для использования в испытательных лабораториях при проведении сертификационных испытаний средств криптографической защиты информации на соответствие требованиям СТБ 34.101.27-2022 «Информационные технологии и безопасность. Средства криптографической защиты информации. Требования безопасности».

Содержание

1 Термины, обозначения и сокращения

В настоящем документе применяются следующие термины и сокращения:

выборка	последовательность наблюдений, вырабатываемая источником случайности;
источник случайности	источник случайных данных, представляющих собой целые числа;
коллизия	случай повторения значений выборки в наборе данных;
наблюдение	отдельное числовое значение, получаемое на выходе источника случайности;
н.о.р.	независимые и одинаково распределённые (случайные величины).
энтропия	мера неопределённости (непредсказуемости) данных.

В математической статистике используются различные способы определения энтропии как меры неопределённости данных. В настоящей методике используется минимальная энтропия — минимальная из всех характеристик неопределённости.

В определении энтропии используются логарифмы, основание которых можно выбирать различными способами. В настоящей методике логарифмы берутся по основанию 2, т.е. используется двоичная энтропия.

Под энтропией источника случайности понимается степень неопределённости отдельных наблюдений, а не всей выходной последовательности наблюдений. Другими словами, подразумевается удельная энтропия.

Данная методика основана на документе [1].

2 Объект и цель испытаний

Объектом испытаний является источник случайности, используемый для выработки случайной числовой последовательности.

В компьютерных системах могут использоваться, например, следующие источники случайности:

— физические источники, использующие процессы в физических устройствах (например, шум в радиоэлектронных приборах);

— системные источники, использующие состояния, процессы и события операционной системы (например, системное время, сетевая активность, прерывания);

— источники, основанные на активности операторов (например, движения мышью, нажатия клавиш).

Выходными данными источника случайности является числовая последовательность наблюдений — неотрицательных целых чисел.

На испытания представляются:

1) документация, описывающая источник случайности, его режимы работы, способ формирования наблюдений, множество выходных значений;

2) выходные числовые последовательности, сгенерированные в различных режимах работы источника случайности.

Выходные последовательности могут быть получены независимо или с участием разработчика. Если предполагается независимое снятие выходной последовательности, то разработчик дополнительно должен предоставить необходимое аппаратное и программное обеспечение.

Целью испытаний является оценка энтропии выходной последовательности источника случайности.

Данная методика может использоваться в дополнение к методике испытаний программных средств криптографической защиты информации на соответствие требованиям СТБ 34.101.27 - 2022 в части оценки энтропии источников случайности.

3 Требования к документации

В документации должны быть описаны источник случайности, способ формирования выходных наблюдений, множество возможных выходных значений, размер каждого наблюдения. Дополнительно должны быть отражены недостижимые на практике выходные значения.

При наличии различных режимов работы источника случайности данная информация должна быть предоставлена по каждому режиму.

4 Средства и порядок испытаний

При испытаниях используются следующие методы:

1) проверка выходных последовательностей на независимость и одинаковую распределённость;

2) оценка энтропии выходных последовательностей.

4.1 Проверка выходных последовательностей на независимость и одинаковую распределённость

Проверка выходных последовательностей на н.о.р. состоит в применении статистических тестов, описанных в 5.1, для проверки гипотезы о том, что выходная последовательность является последовательностью н.о.р. случайных величин.

Если для некоторых тестов недостаточно данных и нет возможности получить для проверки выходную последовательность большей длины, то данные тесты пропускаются и их результаты не учитываются.

Гипотеза о том, что выходная последовательность источника случайности является последовательностью н.о.р. случайных величин, принимается, если все исследуемые выходные последовательности проходят все тесты. Если какая-либо выходная последовательность не проходит хотя бы один из тестов, то данная гипотеза отвергается.

При наличии различных режимов работы источника случайности гипотеза о н.о.р. проверяется отдельно для каждого режима работы. При этом проверяются только выходные последовательность, полученные в данном режиме работы.

Дальнейшее тестирование осуществляются согласно принятой по результатам проверки гипотезе.

4.2 Оценка энтропии выходных последовательностей

Оценка энтропии выходных последовательностей состоит в применении специальных тестов для оценки минимальной энтропии. Тесты описаны в подразделе 5.2.

Оценка энтропии осуществляется по-разному для выходных последовательностей, являющихся последовательностью н.о.р. случайных величин и не являющихся таковыми.

Если для некоторых тестов недостаточно данных и нет возможности получить для проверки выходную последовательность большей длины, то данные тесты пропускаются и их результаты не учитываются.

Оценка минимальной энтропии выходной последовательности равна наименьшей из всех оценок минимальной энтропии, полученных различными тестами. Оценка минимальной энтропии источника случайности равна наименьшей оценке минимальной энтропии среди всех исследуемых выходных последовательностей.

При наличии различных режимов работы источника случайности оценка минимальной энтропии производится отдельно для каждого режима работы. При этом учитываются только выходные последовательности, полученные в данном режиме работы.

5 Методы испытаний

5.1 Проверка выходной последовательности на независимость и одинаковую распределённость

В данном разделе описаны тесты, применяемые для проверки того, что анализируемая выборка является последовательностью независимых и одинаково распределённых (н.о.р.) наблюдений. Набор тестов включает в себя тесты перестановок 5.1.1, и хи-квадрат тесты 5.1.2. Если все тесты не отвергают утверждения о н.о.р. (т.е. выборка проходит тестирование), то оценка энтропии выполняется согласно 5.2.1 в предположении, что данные являются н.о.р. Если хотя бы один из тестов отвергает гипотезу о н.о.р., то оценка энтропии выполняется согласно 5.2.2 для данных, не являющихся н.о.р.

5.1.1 Тесты перестановок для проверки независимости и одинаковой распределённости

Если верна нулевая гипотеза о том, что наблюдения в выборке являются н.о.р., то после любой перестановки наблюдений в выборке они остаются независимыми и сохраняют своё распределение. Соответственно, новая выборка, полученная после перестановки наблюдений, должна иметь те же статистические свойства, что и исходная выборка. При этом статистики, вычисленные для исходной выборки и для новой выборки, будут иметь одинаковое распределение вероятностей. Однако, если нулевая гипотеза не выполняется (т.е. наблюдения не являются н.о.р.), то некоторые статистики могут значимо различаться для исходной и новой выборок.

Перед применением тестов перестановок выборка наблюдений разбивается на десять непересекающихся подвыборок одинакового размера. Например, для выборки длины N каждая подвыборка будет иметь длину $\lfloor \frac{N}{10} \rfloor$. Тесты перестановок применяются отдельно для каждой из подвыборок, проверяя н.о.р. её наблюдений.

В основе тестов перестановок лежит вычисление ряда статистических оценок. Предполагается, что относительно малые или относительно большие значения оценок будут свидетельствовать об отклонении выборки от ожидаемой модели н.о.р. случайных величин. В тестах перестановок предлагается использовать следующий набор оценок:

— оценка сжатия 5.1.1.1;

— оценки серий 5.1.1.2;

— оценка максимального кумулятивного отклонения 5.1.1.3;

— оценки направленных серий 5.1.1.4;

— оценка ковариации 5.1.1.5;

— оценки коллизий 5.1.1.6.

Схема тестов перестановок описывается следующим образом.

1) Для каждой из 10 подвыборок производятся следующие действия.

а) Вычисляются все оценки для исходной подвыборки, обозначим их $S_{0,j}^{(k)}$, $1 \leqslant j \leqslant J$ — номер оценки, J — общее число различных оценок, k – номер подвыборки.

б) Для i от 1 до 1000 повторяются следующие шаги:

— осуществляется перестановка наблюдений подвыборки, используя алгоритм Фишера-Йейтса [2];

— вычисляются все оценки $S_{i,j}^{(k)}$ для новой подвыборки, полученной после перестановки.

2) Для каждой подвыборки k и для каждого вида оценки j формируется массив $\Sigma_j^{(k)} = \left\{S_{0,j}^{(k)},\ S_{1,j}^{(k)},\ \ldots,\ S_{1000,j}^{(k)}\right\}$ из 1001 значения оценок и определяется ранг $R_j^{(k)}$, $0 \leqslant R_j^{(k)} \leqslant 1000$, согласно номеру оценки $S_{0,j}^{(k)}$ в отсортированном массиве $\Sigma_j^{(k)}$.

Например, если оценка $S_{0,j}^{(k)}$ будет самой низкой ($S_{0,j}^{(k)} < S_{i,j}^{(k)}$,$1 \leqslant i \leqslant 1000$), то её ранг будет равен 0. Напротив, если она будет самой высокой ($S_{0,j}^{(k)} > S_{i,j}^{(k)}$,$1 \leqslant i \leqslant 1000$), то её ранг будет равен 1000.

В случае совпадения оценки $S_{0,j}^{(k)}$ с одной или несколькими оценками из множества $\Sigma_j^{(k)}$, то в качестве ранга $R_j^{(k)}$ выбирается значение, наиболее близкое к 500.

3) Всего будет получено $10 \times J$ значений рангов. Вероятность, что значение ранга меньше 50 или больше 950, составляет 10%. Для каждого вида оценки j, $1 \leqslant j \leqslant J$, во множестве рангов $\left\{R_j^{(1)}, R_j^{(2)}, \ldots, R_j^{(10)}\right\}$ определяется число рангов, меньших 50 или больших 950.

Если для какой-то оценки j число таких рангов будет равно восьми или более, то выборка не проходит проверку и полагается, что её наблюдения не являются н.о.р.

5.1.1.1 Оценка сжатия

Алгоритмы сжатия предназначены для устранения избыточности в строке символов, связанной с частым повторением некоторых последовательностей символов. Оценка сжатия выборки данных является мерой качества сжатия и определяется как длина сжатой последовательности. Если наблюдения исходной выборки не являются н.о.р., то в ней могут присутствовать часто повторяющиеся шаблоны (группы наблюдений), которые должны исчезнуть после перестановки наблюдений. Соответственно, оценка сжатия может отличаться для исходной выборки и выборки, полученной после перестановки наблюдений.

Оценка сжатия строится следующим образом.

1) Наблюдения из выборки данных кодируются в виде символьной строки, содержащей список значений, разделенных запятыми: например, «144,21,139,0,0,15».

2) Символьная строка обрабатывается алгоритмом сжатия BZ2 [3].

3) Оценка сжатия будет равна длине сжатой строки в байтах.

5.1.1.2 Оценки серий

Если наблюдения выборки н.о.р., то максимальная длина серии наблюдений, больших или меньших медианы, должна быть не слишком большой и не слишком малой. Аналогично, число серий наблюдений, больших или меньших медианы, также не должно быть

слишком большим или слишком малым. Слишком большое или слишком малое значение максимальной длины серий или числа серий свидетельствует о наличии зависимости наблюдений выборки.

Оценки серий строятся следующим образом.

1) Вычисляется медиана выборки. Если выборка бинарная, то медиана равна 0,5.

2) Строится вспомогательная подвыборка на основе наблюдений исследуемой выборки по следующим правилам:

а) если наблюдение больше медианы, то во вспомогательную подвыборку добавляется значение «+1»;

б) если наблюдение меньше медианы, то во вспомогательную подвыборку добавляется значение «–1»;

в) если наблюдение совпадает с медианой, то во вспомогательную подвыборку ничего не добавляется, т.е. значения, равные медиане, игнорируются.

3) Определяется наиболее длинная серия, состоящая только из «+1» или только из «–1» во вспомогательной подвыборке. Длина этой серии является первой оценкой.

4) Второй оценкой будет число серий, состоящих только из «+1» или только из «–1».

Пример 5.1. *Предположим, что исходная выборка состоит из семи наблюдений {5, 15, 12, 1, 13, 9, 4}. Медиана данной выборки равна 9.*

Будет построена следующая вспомогательная подвыборка: {–1, +1, +1, –1, +1, –1}.

Во вспомогательной подвыборке можно выделить следующие серии: (–1), (+1, +1), (–1), (+1) и (–1).

Первая оценка, максимальная длина серии, равна 2. Вторая оценка, общее число серий, равна 5.

5.1.1.3 Оценка максимального кумулятивного отклонения

Оценка кумулятивного отклонения позволяет определить, образуются ли в выборке кластеры наблюдений с относительно большими или малыми значениями. Если наблюдения выборки н.о.р., то кумулятивное отклонение не должно быть слишком большим или слишком малым. Малое значение оценки кумулятивного отклонения показывает наличие некоторого процесса, предотвращающего даже случайное появление кластеров наблюдений с относительно большими или малыми значениями. Большое значение оценки кумулятивного отклонения показывает слишком частое появление таких кластеров.

Оценка максимального кумулятивного отклонения является мерой отклонения кумулятивной суммы значений наблюдений от её математического ожидания. Пусть выборка длины N состоит из наблюдений $\{s_1, s_2, \ldots s_N\}$ и среднее значение наблюдений равно μ, тогда кумулятивное отклонение в точке i будет равно $d_i = s_1 + s_2 + \ldots + s_i - i\mu$, $1 \leqslant i \leqslant N$. Оценкой является максимальное значение кумулятивного отклонения d_i.

Оценка максимального кумулятивного отклонения строится следующим образом.

1) Вычисляется выборочное среднее μ.

2) Для $i = 1, 2, \ldots, N$ вычисляются величины кумулятивных отклонений:

$$d_i = \left| \sum_{j=1}^{i} s_j - i\mu \right|.$$

3) Оценка максимального кумулятивного отклонения будет равна $d_{\max} = \max\limits_i d_i$.

Пример 5.2. *Пусть выборка состоит из пяти наблюдений* $\{2, 15, 4, 10, 9\}$ *со средним значением* $\mu = 8$. *Кумулятивные отклонения имеют следующие значения:*

$$d_1 = |2 - 8| = 6;$$
$$d_2 = |(2 + 15) - 2 \cdot 8| = 1;$$
$$d_3 = |(2 + 15 + 4) - 3 \cdot 8| = 3;$$
$$d_4 = |(2 + 15 + 4 + 10) - 4 \cdot 8| = 1;$$
$$d_5 = |(2 + 15 + 4 + 10 + 9) - 5 \cdot 8| = 0.$$

Таким образом, оценка максимального кумулятивного отклонения $d_{\max} = 6$.

5.1.1.4 Оценки направленных серий

Для построения оценок направленных серий исследуются возрастающие и убывающие серии наблюдений. Слишком большое или малое значение числа серий и максимальной длины серии возрастающих (убывающих) наблюдений свидетельствует о наличии зависимости наблюдений выборки.

Пусть выборка состоит из наблюдений $\{s_1, s_2, \ldots, s_N\}$. Если выборка является бинарной, то необходима предварительная обработка, в результате которой биты объединяются в байты и в качестве новых наблюдений выступают веса Хэмминга — число единиц в байте.

Оценки направленных серий строятся следующим образом.

1) По исходной выборке строится вспомогательная выборка по следующим правилам:

а) последовательно рассматриваются пары (s_1, s_2), (s_2, s_3), ... , (s_{N-1}, s_N);

б) если в паре (s_i, s_{i+1}) значение первого наблюдения больше второго, то к вспомогательной выборке добавляется элемент «–1»;

в) если в паре (s_i, s_{i+1}) значение первого наблюдения меньше второго, то к вспомогательной выборке добавляется элемент «+1»;

г) если в паре (s_i, s_{i+1}) значения наблюдений равны, то к вспомогательной выборке добавляется элемент «0».

2) Из вспомогательной выборки удаляются начальные нули.

Например, вспомогательная выборка $\{0, 0, -1, 0, 0, 1\}$ превратится в $\{-1, 0, 0, 1\}$, а выборка $\{1, 0, 0, -1, 0\}$ по результатам данной операции не изменится.

3) Первой оценкой является общее число серий из элементов «+1» («–1»), при этом элемент «0» не прерывает серию.

Например, серия (+1, 0, +1, +1, 0, 0, +1, ...) будет являться серией из элементов «+1».

4) Второй оценкой является максимальная длина серии из элементов «+1» («–1»), при этом элемент «0» не прерывает серию.

5) Подсчитывается число элементов «+1» и число элементов «–1». Третьей оценкой является максимальное из этих значений.

Пример 5.3. *Пусть выборка состоит из наблюдений* $\{2, 2, 2, 5, 7, 7, 9, 3, 1, 4, 4\}$. *Вспомогательная выборка будет иметь вид* $\{0, 0, +1, +1, 0, +1, -1, -1, +1, 0\}$. *Можно выделить следующие серии, отбрасывая начальные нули из вспомогательной выборки: (+1, +1, 0, +1), (–1, –1) и (+1, 0).*

Первая оценка, число серий, будет равна 3. Вторая оценка, максимальная длина серии, будет равна 4. Число элементов «+1» равно 4, число элементов «–1» равно 2. Третья оценка, максимум из этих значений, будет равна 4.

Пример 5.4. *Рассмотрим бинарную выборку, которая преобразуется в следующую последовательность байтов, записанных в шестнадцатеричном виде:* $\{A3, 57, 3F, 42, BD\}$. *По весам Хэмминга* $\{4, 5, 6, 2, 6\}$ *строим вспомогательную выборку* $\{+1, +1, -1, +1\}$, *в которой можно выделить следующие серии: (+1, +1), (–1) и (+1).*

Первая оценка, число серий, будет равна 3. Вторая оценка, максимальная длина серии, будет равна 2. Число элементов «+1» равно 3, число элементов «–1» равно 1. Третья оценка, максимум из этих значений, будет равна 3.

5.1.1.5 Оценка ковариации

Оценка ковариации является мерой связи между соседними наблюдениями. При наличии линейной зависимости между соседними наблюдениями ковариация для исходной выборки должна быть по модулю больше, чем ковариация новой выборки, полученной перестановкой наблюдений.

Пусть выборка длины N состоит из наблюдений $\{s_1, s_2, \ldots, s_N\}$, и среднее значение равно μ. Оценка ковариации вычисляется по следующей формуле:

$$\frac{1}{N-1}\sum_{i=1}^{N-1}(s_i - \mu)(s_{i+1} - \mu).$$

Пример 5.5. *Пусть выборка состоит из пяти наблюдений* $\{15, 2, 6, 10, 12\}$ *со средним значением* $\mu = 9$.

Оценка ковариации равна

$$\frac{1}{4}\left[(15-9)(2-9) + (2-9)(6-9) + (6-9)(10-9) + (10-9)(12-9)\right] = \\ = \frac{1}{4}\left[-42 + 21 - 3 + 3\right] = -\frac{21}{4}.$$

5.1.1.6 Оценки коллизий

Оценки коллизий являются мерой числа наблюдений до появления повторного значения (коллизии). Предполагается, что если в выборке вероятности значений меняются со временем, то в исходной выборке потребуется меньше наблюдений до появления коллизии, чем в её перестановках.

Пусть выборка состоит из наблюдений $\{s_1, s_2, \ldots, s_N\}$. Если выборка является бинарной, то необходима предварительная обработка, в результате которой биты объединяются в байты.

Оценки коллизий строятся следующим образом.

1) Для подсчёта числа наблюдений до коллизии используется список C, который инициализируется пустым множеством.

2) Устанавливается текущая позиция $p = 1$.

3) Пока $p < N$, выполнять следующие действия:

а) найти минимальное i, что последовательность $s_p, s_{p+1}, \ldots, s_{p+i}$ содержит ровно одно повторяющееся наблюдение; если такого i не существует, то перейти к шагу 4;

б) добавить i в список C;

в) сдвинуть текущую позицию $p \leftarrow p + i + 1$.

4) Оценками коллизий будут являться минимальное значение, среднее всех значений и максимальное значение в списке C.

Пример 5.6. *Пусть выборка состоит из наблюдений* $\{2,1,1,2,0,1,0,1,1,2\}$. *Если первое наблюдение «2» находится на позиции 1, то первая коллизия наблюдается при* $i=2$. *Текущая позиция сдвигается на 3 наблюдения, после чего непросмотренная часть выборки имеет вид* $\{2,0,1,0,1,1,2\}$. *Следующая коллизия наблюдается при* $i=3$. *Текущая позиция сдвигается на 4 наблюдения, непросмотренная часть выборки имеет вид* $\{1,1,2\}$. *Следующая коллизия наблюдается при* $i=1$. *Текущая позиция сдвигается на 2 наблюдения, оставшаяся часть выборки имеет вид* $\{2\}$, *и больше невозможно обнаружить коллизии.*

Построен список $C=\{2,3,1\}$. *Первая оценка, минимальное значение в списке, будет равна 1. Вторая оценка, среднее всех значений в списке, будет равна 2. Третья оценка, максимальное значение в списке, будет равна 3.*

5.1.2 Хи-квадрат тесты для проверки независимости и одинаковой распределённости

Если наблюдения выборки н.о.р., то её можно рассматривать как выборку независимых величин с мультиномиальным распределением. Для проверки соответствия выборки наблюдений мультиномиальному распределению используется критерий хи-квадрат.

Если множество значений наблюдений слишком велико, то вероятность отдельных значений становится низкой, следовательно, для применения хи-квадрат теста требуются большие объёмы данных. Если необходимое число данных собрать невозможно, то хи-квадрат тест не выполняется.

Используется два различных типа хи-квадрат критериев: критерий независимости и критерий согласия. Хи-квадрат критерий независимости предназначен для обнаружения зависимости в вероятностях между соседними наблюдениями и описан в 5.1.2.1 для небинарных данных и в 5.1.2.3 для бинарных данных. Критерий согласия хи-квадрат предназначен для проверки согласия распределения подвыборок, сформированных из исходной выборки и описан в 5.1.2.2 для небинарных данных и в 5.1.2.4 для бинарных данных.

5.1.2.1 Проверка независимости для небинарных данных

В данной проверке на первом шаге по выборке оцениваются вероятности всех возможных значений наблюдений. На втором шаге по выборке оцениваются вероятности соседних пар значений и проверяется согласие распределения вероятностей пар значений с теоретическим.

Пусть $p(x_i)$ — оценка вероятности того, что значение наблюдения равно x_i, $1 \leqslant i \leqslant n$, n — число различных значений наблюдений. Обозначим $E(x_i, x_j)$ ожидаемое число появлений пары (x_i, x_j) в анализируемой выборке. Пусть список *List1* содержит все пары (x_i, x_j), для которых $E(x_i, x_j) \geqslant 5$, и *List2* содержит все остальные пары (x_i, x_j). Обозначим E_0 ожидаемое число появления в выборке всех пар из списка *List2*.

Проверка на независимость для небинарных данных выполняется следующим образом.

1) По выборке оцениваются вероятности $p(x_i)$ всех возможных значений наблюдений x_i по следующей формуле:

$$p(x_i) = \frac{n(x_i)}{N},$$

где $n(x_i)$ определяет, сколько раз в выборке встретилось наблюдение x_i, а N — длина выборки.

2) Определяется максимальная вероятность $p_{\max} = \max\limits_i p(x_i)$.

3) Подсчитывается число параметров $q = 1 + \sum\limits_i I(p(x_i)p_{\max} \geqslant 5/N)$, где $I(\cdot)$ — индикаторная функция.

Если $q = 1$, то объём выборки слишком мал и алгоритм прекращает работу. Результат проверки в этом случае — «недостаточно данных».

4) Списки *List1* и *List2* инициализируются пустыми множествами, и устанавливается начальное значение частоты $E_0 = 0$.

5) Для каждой возможной пары значений (x_i, x_j), включая (x_i, x_i), выполняются следующие операции:

а) вычисляется значение $E(x_i, x_j) = p(x_i)p(x_j)(N-1)$;

б) если $E(x_i, x_j) \geqslant 5$, то пара (x_i, x_j) добавляется в список *List1*, иначе пара (x_i, x_j) добавляется в список *List2* с пересчётом значения $E_0 \leftarrow E_0 + E(x_i, x_j)$;

6) Проверяется наличие достаточного числа степеней свободы. Обозначим w число пар в списке *List1*. Если $w + 1 - q < 1$, то объём выборки слишком мал и алгоритм прекращает работу. Результат проверки в этом случае — «недостаточно данных».

7) Вычисляется хи-квадрат статистика χ^2. Обозначим $n(x_i, x_j)$ наблюдаемую частоту пары (x_i, x_j) в выборке. Выполняются следующие вычисления:

а) инициализируется значение статистики $\chi^2 = 0$;

б) для каждой пары (x_i, x_j) из списка *List1* пересчитывается значение статистики χ^2 следующим образом:

$$\chi^2 \leftarrow \chi^2 + \frac{(E(x_i, x_j) - n(x_i, x_j))^2}{E(x_i, x_j)};$$

в) инициализируется наблюдаемая частота пар из списка *List2* $n_0 = 0$;

г) для каждой пары (x_i, x_j) из списка *List2* пересчитывается наблюдаемая частота:

$$n_0 \leftarrow n_0 + n(x_i, x_j);$$

д) пересчитывается значение статистики χ^2 следующим образом:

$$\chi^2 \leftarrow \chi^2 + \frac{(E_0 - n_0)^2}{E_0}.$$

8) Статистика χ^2 сравнивается с пороговым значением $\chi^2_{w+1-q}(\alpha)$ распределения хи-квадрат с $w + 1 - q$ степенями свободы и уровнем значимости $\alpha = 0,001$.

Если $\chi^2 > \chi^2_{w+1-q}(\alpha)$, то выборка не проходит проверку на независимость наблюдений.

5.1.2.2 Проверка одинаковой распределённости для небинарных данных

В данной проверке анализируемая выборка разбивается на подвыборки, а затем проверяется гипотеза о согласии распределения каждой подвыборки с распределением исходной выборки. Если гипотеза о согласии отклоняется, то это свидетельствует о неоднородности выборки.

Проверка одинаковой распределённости для небинарных данных осуществляется следующим образом.

1) Исходная выборка длины N разбивается на 10 подвыборок длины $\lfloor N/10 \rfloor$.

2) Для каждого возможного значения наблюдения x_i, $1 \leqslant i \leqslant n$, где n — число различных значений наблюдений, по исходной выборке оценивается ожидаемая частота $E(x_i)$

его появления в подвыборке:

$$E(x_i) = \frac{n(x_i)}{N} \cdot \left\lfloor \frac{N}{10} \right\rfloor,$$

где $n(x_i)$ определяет, сколько раз в исходной выборке встретилось наблюдение x_i.

3) Создаётся список *List3* всех возможных значений наблюдений x_i таких, что $E(x_i) \geqslant 5$.

4) Создаётся список *List4* всех возможных значений наблюдений x_i таких, что $E(x_i) < 5$.

5) Вычисляется ожидаемая суммарная частота E_0 встречаемости всех значений наблюдений x_i из списка *List4*, как сумма соответствующих значений частот $E(x_i)$.

6) Для каждой из десяти подвыборок ($1 \leqslant j \leqslant 10$) повторяются следующие операции:

а) устанавливается начальное значение статистики $\chi_j^2 = 0$;

б) для каждого значения x_i из списка *List3* пересчитывается значение статистики χ_j^2:

$$\chi_j^2 \leftarrow \chi_j^2 + \frac{(E(x_i) - n_j(x_i))^2}{E(x_i)},$$

где $n_j(x_i)$ определяет, сколько раз в j-ой подвыборке встретилось наблюдение x_i;

в) определяется величина $n_{0,j}$, сколько раз в j-ой подвыборке встретились наблюдения x_i из списка *List4*, как сумма соответствующих наблюдаемых значений $n_j(x_i)$;

г) пересчитывается значение статистики χ_j^2:

$$\chi_j^2 \leftarrow \chi_j^2 + \frac{(E_0 - n_0)^2}{E_0}.$$

7) Статистики χ_j^2, $1 \leqslant j \leqslant 10$, сравниваются с пороговым значением $\chi_L^2(\alpha)$ распределения хи-квадрат с L степенями свободы, где L — размер списка *List*3, и уровнем значимости $\alpha = 0,001$.

Если $\chi_j^2 > \chi_L^2(\alpha)$ для некоторой подвыборки j, $1 \leqslant j \leqslant 10$, это означает, что распределение данной подвыборки не согласуется с распределением всей выборки. В этом случае анализируемая выборка не проходит проверку на одинаковую распределённость наблюдений.

5.1.2.3 Проверка независимости для бинарных данных

Особенностью бинарных данных является то, что наблюдения принимают лишь два различных значения, кодируемых битом 0 или 1. Если наблюдения бинарной выборки н.о.р., то распределения любых двух битов выборки будет одинаковыми, следовательно, вероятность любой k-битной строки будет равна произведению вероятностей значений составляющих её битов.

Данный тест проверяет, согласуется ли распределение k-битных строк с ожидаемым распределением. Если расположенные рядом биты не являются независимыми, то вероятности k-битных строк не будут равны произведению вероятностей составляющих их битов и гипотеза о согласии распределений будет отвергнута.

Проверка независимости для бинарных данных осуществляется следующим образом.

1) Определяются следующие величины: C_0 — число нулевых битов в анализируемой выборке, C_1 — число единичных битов, $C_x = \min(C_0, C_1)$, $p = C_1/N$ — частота единичных битов, N — общее число битов в выборке.

2) Для k от 2 до 11, пока выполняется условие $(C_x/N)^k > 5/N$, осуществляются следующие действия:

а) строится модифицированная выборка, объединяя последовательно каждые k битов в одно наблюдение со значением v, $0 \leqslant v \leqslant 2^k - 1$, при этом вероятность значения v составит

$$p_v = \frac{N}{k} p^{W(v)} (1-p)^{k-W(v)},$$

где вес Хэмминга $W(v)$ определяет число единичных битов в v;

б) вычисляется значение статистики χ^2:

$$\chi^2 = \sum_{v=0}^{2^k-1} \frac{(n_v - p_v)^2}{p_v},$$

где n_v — число наблюдений со значением v в модифицированной выборке.

3) Статистика χ^2 сравнивается с пороговым значением $\chi^2_{2^k-1}(\alpha)$ распределения хи-квадрат с $2^k - 1$ степенями свободы, и уровнем значимости $\alpha = 0,001$.

Если $\chi^2 > \chi^2_{2^k-1}(\alpha)$, то выборка не проходит проверку на независимость.

5.1.2.4 Проверка одинаковой распределённости для бинарных данных

В данной проверке анализируемая выборка разбивается на подвыборки, а затем исследуется гипотеза о согласии распределений каждой подвыборки с распределением исходной выборки. Если гипотеза о согласии отклоняется, это свидетельствует о неоднородности выборки.

Проверка одинаковой распределённости для бинарных данных осуществляется следующим образом.

1) Подсчитывается число единичных битов в исходной выборке, равное n, и оценивается вероятность появления единичного бита $p = n/N$, где N — общее число битов в выборке.

2) Исходная выборка длины N разбивается на 10 подвыборок длины $N' = \lfloor N/10 \rfloor$.

3) Вычисляется значение статистики χ^2:

$$\chi^2 = \sum_{i=1}^{10} \frac{(n_i - pN')^2}{pN'},$$

где i определяет номер подвыборки, $1 \leqslant i \leqslant 10$, а n_i — число единиц в i-ой подвыборке.

4) Статистика χ^2 сравнивается с пороговым значением $\chi^2_9(\alpha)$ распределения хи-квадрат с 9 степенями свободы, и уровнем значимости $\alpha = 0,001$.

Если $\chi^2 > \chi^2_9(\alpha) = 27.9$, то выборка не проходит проверку на одинаковую распределённость.

5.2 Оценка минимальной энтропии выходной последовательности

Оценка энтропии для последовательности н.о.р. наблюдений осуществляется согласно пункту 5.2.1. Оценка энтропии для последовательности наблюдений, не являющихся н.о.р., производится согласно пункту 5.2.2.

5.2.1 Оценка минимальной энтропии последовательности независимых и одинаково распределённых случайных величин

Оценка минимальной энтропии последовательности н.о.р. случайных величин производится на основе частоты наиболее встречаемого значения. Чтобы не переоценить значение энтропии, используется верхняя граница 99,0% ого доверительного интервала в качестве оценки вероятности наиболее встречаемого значения.

Пусть выборка длины n состоит из наблюдений $\{s_1, s_2, \ldots, s_n\}, s_i \in A$. Мощность множества (алфавита) A равна $k(k \geqslant 2)$.

Алгоритм вычисления энтропии состоит из следующих шагов.

1) Определить наиболее встречаемое в выборке значение. Пусть оно встретилось n_{max} раз.

2) Вычислить верхнюю границу 99,0%-ого доверительного интервала для вероятности наиболее встречаемого значения:

$$p_u = p + 2.576\sqrt{\frac{p(1-p)}{n-1}},$$

где $p = n_{max}/n$ — оценка вероятности наиболее встречаемого значения, а число 2.576 — квантиль нормального распределения для доверительной вероятности $0.995(u_{0.995})$.

3) Оценка энтропии:

$$H_{min} = -\log_2(p_u).$$

Пример 5.7. *Пусть выборка длины $n = 20$ состоит из наблюдений {0,1,1,2,0,1,2,2,0,1,0,1,1,0,2,2,1,0,2,1}. Определим частоты всех значений (0, 1 и 2): $n_0 = 6$, $n_1 = 8$, $n_2 = 6$. Таким образом, самым встречаемым значением будет 1 и $n_{\max} = n_1 = 8$. Оценка его вероятности будет равна $p = 8/20 = 0.4$. Верхняя граница 99,0%-ого доверительного интервала равна:*

$$p_u = 0.4 + 2.576\sqrt{\frac{0.4 \cdot (1-0.4)}{19}} = 0.6895.$$

Оценка энтропии равна:

$$H_{min} = -\log_2(p_u) = -\log_2(0.6895) = 0.5363.$$

5.2.2 Оценка минимальной энтропии последовательности случайных величин, не являющихся независимыми и одинаково распределёнными

Для оценки энтропии последовательности случайных величин, не являющихся н.о.р, используются статистические тесты 5.2.2.1 – 5.2.2.10. Во всех тестах используется доверительная вероятность $99,0\%$, что соответствует уровню значимости $\alpha = 0.005$.

В качестве итоговой оценки минимальной энтропии используется минимальная из всех оценок, полученных каждым тестом.

Если для каких-то из представленных здесь тестов недостаточно данных и нет возможности получить для проверки более длинную выходную последовательность, то данные тесты пропускаются и их результаты не учитываются при определении оценки минимальной энтропии.

Общие параметры тестов: $n(n > 1)$ — длина выборки $S = \{s_1, s_2, \ldots, s_n\}$; наблюдаемые значения из некоторого алфавита: $s_i \in A = \{x_1, \ldots, x_k\}$, где $k(k \geqslant 2)$ — мощность алфавита.

Если в тестах дополнительно используются другие параметры, то их описание приводится непосредственно в тестах.

Три теста следует применять только к бинарным последовательностям, это тесты: коллизий (п. 5.2.2.2), Маркова (п. 5.2.2.3) и сжатия (п. 5.2.2.4).

5.2.2.1 Частотный тест

Общие сведения. Частотная статистика вычисляет энтропию на основе относительной частоты $p_{\max}$ наиболее распространенного значения в выборке. Минимальная энтропия вычисляется на основе верхней границы доверительного интервала оценки относительной частоты p_u. Тест даст консервативную оценку энтропии на основе любого количества собранных данных.

Краткое описание реализации. Исходными данными является массив или последовательность наблюдений из источника случайности. Осуществляется проход по массиву с подсчетом частоты встречаемости каждого значения выборки. На основе вычисленных частот определяется наиболее частое значение выборки, а затем вычисляется верхняя граница доверительного интервала для вероятности наиболее частого значения. Минимальная энтропия вычисляется как отрицательный логарифм верхней границы вероятности наиболее встречаемого значения выборки.

Алгоритм частотного теста состоит из следующих шагов.

1) Вычислить относительную частоту p наиболее частого значения:

$$p = \max_i \frac{\#\{x_i \in S\}}{n}.$$

2) Вычислить верхнюю границу $99,0\%$ ого доверительного интервала для вероятности наиболее встречаемого значения:

$$p_u = p + 2.576\sqrt{\frac{p(1-p)}{n-1}},$$

где число 2.576 — квантиль нормального распределения $u_{0.995}$ для доверительной вероятности 99.0%.

3) Оценки энтропии равна:

$$H_{\min} = -\log_2(p_u).$$

Замечание 5.1. *Заметим, что вычисления для частотного теста схожи с шагами вычисления вычислениями минимальной энтропии для последовательностей IID.*

Условия применения. Данный тест не требует больших объемов данных, тем не менее, рекомендуемая длина выборки $n \approx 1\ 000\ 000$.

5.2.2.2 Тест коллизий

В тесте коллизий измеряется среднее время до первого совпадения (коллизии) в массиве наблюдений. Данный тест даёт низкую оценку энтропии для источников случайности, которые имеют значительное смещение частоты нескольких значений от остальных, что приводит к более быстрому появлению коллизий.

Тест коллизий находит нижнюю границу энтропии с заданным уровнем доверия в случае независимых наблюдений. Наличие зависимостей между наблюдениями может привести к завышенной оценке энтропии по результатам данного теста.

Пусть выборка длины n состоит из наблюдений $\{s_1, s_2, \ldots, s_n\}$.

Алгоритм теста коллизий состоит из следующих шагов.

1) Инициализировать счётчик числа коллизий $\nu = 0$ и индекс начального момента коллизии $index = 1$.

2) Выполнить поиск коллизии — такого минимального натурального j, что $s_j = s_i$ для некоторого i, $index \leqslant i < j$.

3) Увеличить счётчик числа коллизий: $\nu := \nu + 1$, в последовательность коллизий добавить элемент $t_\nu = j - index + 1$ и увеличить индекс начала поиска следующей коллизии $index = j + 1$.

Время коллизии t_ν определяется разностью соседних значений моментов коллизий.

4) Повторить шаги 2 и 3 пока имеются данные.

5) Вычислить выборочные среднее μ и стандартное отклонение σ длительностей коллизии:

$$\mu = \frac{1}{\nu}\sum_{i=1}^{\nu} t_i, \qquad \sigma = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(t_i - \mu)^2}.$$

6) Вычислить нижнюю границу доверительного интервала для среднего, основанного на нормальном распределении, с доверительным вероятностью $99,0\%$:

$$\bar{\mu} = \mu - 2.576\frac{\sigma}{\sqrt{\nu}},$$

где число 2.576 соответствует значению квантиля нормального распределения $u_{0.995}$.

7) Используя бинарный поиск, решить относительно p:

$$\bar{\mu} = pq^{-2}\left(1 + \frac{p^{-1} - q^{-1}}{2}\right)F(q) - \frac{pq^{-1}(p^{-1} - q^{-1})}{2},$$

где $q = 1 - p, \quad p \geqslant q, \quad F(1/z) = \Gamma(3, z)z^{-3}e^z$, $\Gamma(\cdot, \cdot)$ — неполная гамма-функция:

$$\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt.$$

Границами бинарного поиска является интервал $[0,5; 1]$.

8) Если бинарный поиск дает решение, то оценка минимальной энтропии:

$$H_{min} = -\log_2(p).$$

Замечание 5.2. *Число коллизий зависит от размера выборки и мощности множества выходных значений источника случайности. Если мощность множества выходных значений достаточно велика, то это может накладывать невыполнимые на практике требования на длину выборки.*

Пример 5.8. *Пусть* $S = (1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,1,1,1,0,0,1,1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,1,1,1,0)$. *Длина выборки* $n = 40$. *Последовательность коллизий равна:* $(1,0,0)$, $(0,1,1)$, $(1,0,0)$, $(1,0,1)$, $(0,1,0)$, $(1,1)$, $(1,0,0)$, $(1,1)$, $(0,0)$, $(0,1,1)$, $(1,0,0)$, $(1,0,1)$, $(0,1,0)$, $(1,1)$. *После шага 5,* $\nu = 14$ *и последовательность длин коллизий равна:*

$(t_1, \ldots, T_{(\nu)}) = (3, 3, 3, 3, 3, 2, 3, 2, 2, 3, 3, 3, 3, 2)$. *Выполнив шаги 5 и 6 алгоритма получим* $\mu = 2.7143, \sigma = 0.4688$ *и* $\bar{\mu} = 2.3915$.

Решая уравнение шага 7 относительно p *находим оценку* $p = 0.7329$*, что позволяет вычислить минимальную энтропию:*

$$H_{min} = -\log_2(p) = -\log_2(0.7329) = 0.4483.$$

5.2.2.3 Тест марковской зависимости

В тесте марковской зависимости строится оценка энтропии на основе зависимостей между соседними наблюдениями выходной последовательности. В качестве модели зависимости используется однородная цепь Маркова первого порядка (ОЦМ), в которой значение следующего наблюдения зависит только от значения текущего наблюдения.

По исследуемой выходной последовательности оценивается матрица вероятностей переходов и вектор вероятностей начальных состояний.

Ключевым компонентом в оценке энтропии теста марковской зависимости является способность точно оценить матрицу одношаговых переходных вероятностей ОЦМ). Чем больше данных предоставляется, тем точнее будут оценки вероятностей переходных вероятностей. Тест марковской зависимости оценки энтропии можно применять к произвольным данным, однако ниже рассмотрим его применение только к бинарным входным данным (размер алфавита равен 2).

Пусть выборка длины n состоит из бинарных наблюдений $S = \{s_1, s_2, \ldots, s_n\}, s_i \in A = \{0, 1\}$. Таким образом, число различных состояний цепи Маркова будет равно 2 ($k = 2$).

Алгоритм теста марковской зависимости состоит из следующих шагов.

1) Вычислить оценку вектора вероятностей начальных состояний $\pi = (\pi_0, \pi_1)$:

$$\pi_i = \frac{n_i}{n},$$

где n_i — число наблюдений со значением i ($i \in \{0, 1\}$) в выборке S.

2) Вычислить оценку матрицы вероятностей переходов $P = (p_{ij})$ размерности 2×2:

$$p_{ij} = \frac{n_{ij}}{n_i},$$

где n_{ij} — наблюдаемое в выборке число переходов из состояния i в состояние j, при этом $0 \leqslant i, j \leqslant 1$.

3) Используя оценки вектора π и матрицы P вычислить вероятности получения 6 последовательностей длины 128 бит из таблицы 1:

Таблица 1 — Вероятности появления строк длительности 128 бит

Строка	Вероятность строки
$00\ldots0$	$\pi_0 \times p_{0,0}^{127}$
$0101\ldots01$	$\pi_0 \times p_{0,1}^{64} \times p_{1,0}^{63}$
$011\ldots1$	$\pi_0 \times p_{0,1} \times p_{1,0}^{126}$
$100\ldots0$	$\pi_1 \times p_{1,0} \times p_{0,0}^{126}$
$1010\ldots10$	$\pi_1 \times p_{1,0}^{64} \times p_{0,1}^{63}$
$11\ldots1$	$\pi_1 \times p_{1,1}^{127}$

4) Пусть p_u — наибольшая (максимум) из вероятностей в таблице 1.

Оценка минимальной энтропии равна:

$$H_{\min} = \min(-\log_2(p_u)/128, 1).$$

.

Замечание 5.3. *Для оценки вектора начальных вероятностей и матрицы вероятностей переходов необходимо как минимум* $120(20 \times (2+2 \times 2) = 120)$ *бинарных наблюдений. При этом учитываются только достижимые на практике значения.*

В данном алгоритме методики рассматриваются только бинарные выходные последовательности соответствующие однородной цепи Маркова первого порядка, поэтому на практике требования на длину выборки практически не предъявляются.

Условия применения. Тест предназначен для данных практически любой длины ($n \geqslant 120$), рекомендуемая длина выборки $n \approx 1\ 000\ 000$.

Пример 5.9. *Тестовый пример носит иллюстративный характер.*

Пусть $S = (1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,1,1,1,0,0,1,1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,1,1,1,0)$. *Длина выборки* $n = 40$.

Оценки начальных вероятностей равны: $\pi_0 = 0.475,\ \pi_1 = 0.525$. *Оценки элементов матрицы одношаговых переходов равны:* $p_{00} = 0.389, p_{01} = 0.611, p_{10} = 0.571, p_{11} = 0.429$.

На шаге 3 алгоритма по полученным оценкам вектра начальных вероятностей и матрицы одношаговых переходов сформируем таблицу вероятнотстей появления строк 2.

Таблица 2 — Вероятности появления строк длительности 128 бит

Строка	Вероятность строки
$00\ldots0$	3.9837×10^{-53}
$0101\ldots01$	4.4813×10^{-30}
$011\ldots1$	1.4202×10^{-42}
$100\ldots0$	6.4631×10^{-53}
$1010\ldots10$	4.6288×10^{-30}
$11\ldots1$	1.1021×10^{-47}

Анализируя строки таблицы 2 заметим, что максимальная вероятность находится в пятой строке и равна $p_u = 4.6288 \times 10^{-30}$. *По этой вероятности вычислим оценку минимальной энтропии:*

$$H_{\min} = \min(-\log_2(4.6288 \times 10^{-30})/128, 1) = \min(0.761, 1) = 0.761.$$

5.2.2.4 Тест сжатия

Тест сжатия основан на универсальном тесте Маурера и определяет оценку энтропии на основе меры того, насколько сильно может быть сжата выходная последовательность без потери информации.

Преимуществом теста Маурера является то, что для него не требуется независимость наблюдений. Для вычисления статистики Маурера требуется только один проход по выходной последовательности, что позволяет построить эффективный алгоритм.

Пусть выборка длины n состоит из бинарных наблюдений $\{s_1, s_2, \ldots, s_n\}$. Для выполнения теста сжатия формируется новая выборка из блоков по $b(b = 6)$ бит. Новая выборка разбивается на две непересекающиеся части. Первая часть из $d = 1000$ наблюдений используется в качестве словаря для обучения алгоритма сжатия. Вторая часть наблюдений используется для вычисления статистики Маурера.

bf Алгоритм теста сжатия состоит из следующих шагов.

1) Пусть $b = 6$. Тогда $n1 = \lfloor n/b \rfloor$. Формируем новую b -битную последовательность по непересекающимся отрезкам длины b: $X = \{x_1, x_2, \ldots, x_{n1}\}$. Если n не кратно b, то остаток данных не учитывается.

2) Разделить выборку X на две группы: словарь и тестируемые данные.

а) Создать словарь из первых $d = 1000$ наблюдений $X_{(d)}^{(1)} = (x_1, \ldots, x_d)$.

б) Оставшиеся $\nu = n1 - d$ наблюдений $(x_{d+1}, \ldots, x_{n1})$ будут использоваться для тестирования.

3) Инициализировать словарь `dict` размерности $k = 2^b$ нулями. Для i от 1 до d положить $dict[x_i] = i$. Значение $dict[x_i]$ является индексом последнего вхождения каждого x_i в словарь.

4) Пройти тестовые данные со словарем, созданным шаге 2.

а) Пусть D — массив длины ν.

б) Для i от $d + 1$ до $n1$:

(a) Если $dict[x_i] \neq 0$, тогда $D[i - d] = i - dict[x_i]$. Обновить словарь с индексом самого последнего наблюдения, $dict[s_i] = i$.

(b) Если $dict[x_i] = 0$, то это значение добавить в словарь, т.е. $dict[x_i] = i$. Положить $D[i - d] = i$.

5) Вычислить выборочные среднее μ и среднеквадратическое отклонение σ от $(log_2(D_1), \ldots, log_2(D_\nu))$:

$$\mu = \frac{1}{\nu} \sum \log_2 D_i, \quad c = 0.5907$$

и

$$\sigma = c \sqrt{\frac{1}{\nu - 1} \sum_{i=1}^{\nu} (\log_2 D_i)^2 - \mu^2}.$$

Поправочный коэффициент уменьшает стандартное отклонение, чтобы учесть зависимости в значениях D_i.

6) Вычислить нижнюю граница $99,0\%$ -го доверительного интервала для среднего:

$$\bar{\mu} = \mu - 2.576 \frac{\sigma}{\sqrt{\nu}}.$$

7) Решить относительно p с использованием бинарного поиска так, чтобы следующее равенство было верно:

$$\mu = G(p) + (2^b - 1)G(q),$$

где

$$q = \frac{1-p}{2^b - 1},$$
$$G(z) = \frac{1}{\nu} \sum_{t=d+1}^{n} \sum_{u=1}^{t} \log_2(u) F(z,t,u),$$
$$F(z,t,u) = \begin{cases} z^2(1-z)^{u-1} & \text{если} u < t, \\ z(1-z)^{t-1} & \text{если} u = t. \end{cases}$$

Границы бинарного поиска должны принадлежать интервалу $[2^{-b}; 1]$.

8) Если бинарный поиск дает решение, то оценка минимальной энтропии равна:

$$H_{\min} = -\log_2(p)/b.$$

Если решения не существует, то оценка минимальной энтропии равна:

$$H_{\min} = 1.$$

Пример 5.10. *Пример приведен в иллюстративных целях.*

Пусть $d = 4$ *(вместо 1000),* $L = 48$ *и* $S = (1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,$ $1,1,1,0,0,1,1,0,0,0,1,1,1,0,0,1,0,1,0,1,0,1,1,1,0,1,1,1,0,0,0,1,1)$.

После шага 1 новая последовательность примет вид:

$$X = (100011, 100101, 010111, 001100, 011100, 101010, 111011, 100011).$$

Словарь: $(100011, 100101, 010111, 001100)$ *и последовательность тестирования:* $(011100, 101010, 111011, 100011)$. *Тогда* $\nu = 4$. *Словарь инициализируется на шаге 3 и имеет следующие значения (отображаются только ненулевые значения):* $(1,2,3,4)$.

После шага 4 получим: $D_1 = 5, D_2 = 6, D_3 = 7, D_4 = 7$.

Значения, вычисленные на шаге 5, равны: $\mu = 2.6304$, $\sigma = 0.9074$, *а значение шага 6:* $\hat{\mu} = 1.4617$. *На шаге 7 получаем* $p = 0.5715$, *а на шаге 8 получаем оценку минимальной энтропии —* $H_{\min} = 0,1345$.

5.2.2.5 Тест частичных коллекций

Этот метод анализирует частоту частичных коллекций или t -наборов(пар, троек и т.д.), которые появляются во входном наборе данных, и производит оценку энтропии выборки на основе частоты этих t - наборов. Частота t - наборов $(r_1, r_2, \ldots, r_t)$ в $S = (s_1, \ldots, s_n)$ –– это число таких i, что $s_i = r_1, s_(i+1) = r_2, \ldots, s_(i+t-1) = r_t$. Следует отметить, что наборы могут перекрывать друг друга.

Входные данные: $S = (s_1, \ldots, s_n)$; $s_i \in A = \{x_1, \ldots, x_k\} (k \geqslant 2)$.

Алгоритм теста частичных коллекций (t-набора) состоит из следующих шагов.

1) Найти наибольшее t такое, что число вхождений наиболее частого t-набора в S составляет не менее 35.

2) Пусть $Q[i]$ — количество вхождений наиболее частого i-набора в S для $i = 1, \ldots, t$.

Например, в $S = (2,2,0,1,0,2,0,1,2,1,2,0,1,2,1,0,0,1,0,0,0)$ частоты значений равны: $Q[1] = \max(\#0's, \#1's, \#2's) = \max(9,6,6) = \#0's = 9$, и $Q[2] = 4$ соответствует набору $\{01\} \in S$.

3) Для i от 1 до t вычислить оценку $p_{\max}$:

$$P[i] = Q[i]/(n-i+1), \quad P_{\max}[i] = (P[i])^{1/i}, \quad p_{\max} = \max(P_{\max}[1], \ldots, P_{\max}[t]).$$

4) Вычислить верхнюю границу 99% вероятности наиболее частого значения:

$$p_u = \min(1, p_{\max} + 2.576\sqrt{\frac{p_{\max}(1-p_{\max})}{n-1}}),$$

5) Вычислить оценку минимальной энтропии как:

$$H_{\min} = -log_2(p_u).$$

Условия применения: рекомендуется $n \geqslant 1000000$.

Пример 5.11. *Для целей этого примера предположим, что длина набора равна* 3 *вместо* 35 *на первом шаге.*

Пусть $S = (2,2,0,1,0,2,0,1,2,1,2,0,1,2,1,0,0,1,0,0,0), n = 21$, $Q[1] = \max(\#0's, \#1's, \#2's) = \max(9,6,6) = \#0's = 9$ *и* $Q[2] = 4$ *соответствует набору* $\{01\} \in S$. *Число вхождений наиболее распространенного набора из* 4*-х равно* 2, *что ниже порогового значения* u, *следовательно,* $t = 3$.

На шаге 2 $Q[1] = 9, Q[2] = 4, Q[3] = 3$. *Тогда* $P[1] = 0.4286, P[2] = 0.2, P[3] = 0.1579$. $P_{\max}[1] = 0.4286, P_{\max}[2] = 0.4472, P_{\max}[3] = 0.5405$, *а* $P_{\max} = 0.5405$.

Верхняя граница 99.0% *доверительного интервала:* $p_u = 0.8276$. *Оценка минимальной энтропии составляет* $H_{\min} = -log_2(0.8276) = 0.273$.

5.2.2.6 Тест наибольшей повторяющейся подстроки

Общие сведения. Метод оценивает энтропию коллизий источника, основываясь на количестве повторений подстрок (наборов) внутри набора данных.

Входные данные: $S = (s_1, \ldots, s_n); s_i \in A = x_1, \ldots, x_k (k \geqslant 2)$.

Алгоритм теста наибольшей повторяющейся подстроки (теста LRS**)** состоит из следующих шагов.

1) Найти наименьшее u такое, чтобы число вхождений наиболее частого u-набора в S было не менее 35.

2) Найти наибольшее v такое, чтобы число вхождений наиболее частого v-набора в S составляло, по меньшей мере, 2, а наиболее частого $(v+1)$-набора в S равнялось 1. Другими словами, v — наибольшая длина набора, который повторяется. Если v<u, эта оценка не может быть вычислена.

3) Для W от u до v для W-набора вычислить ожидаемую вероятность коллизии:

$$P_W = \frac{\sum_i C_{C_i}^2}{C_{L-W+1}^2},$$

где C_i^j — число сочетаний из i по j; C_i — число вхождений i-го уникального W-набора.

Вычислить ожидаемую среднюю вероятность столкновения на символ строки: $P_{\max,W} = P_W^{1/W}$.

Пусть $p = \max(P_{\max,u}, \ldots, P_{\max,v})$.

4) Вычислить верхнюю границу 99.0% вероятности наиболее частого значения

$$p_u = \min(1, p + 2.576\sqrt{\frac{p(1-p}{n-1}}.$$

5) Вычислить оценку энтропии:

$$H_{\min} = -log_2(p_u).$$

Условия применения: $n \geqslant 1000000$. Тест используется, когда размеры наборов слишком велики для использования метода t-набора.

Пример 5.12. *Для целей этого примера предположим что отсечка равна 3 вместо 35 на шаге 1.*

Пусть $S = (2,2,0,1,0,2,0,1,2,1,2,0,1,2,1,0,0,1,0,0,0)$ *и* $n = 21$. *На шаге 1:* $u = 4$, *поскольку частота наиболее встречаемого набора из четырех элементов равна 2. На шаге 2:* $v = 5$. *После шага 3:* $P_4 = 0.0131, P_5 = 0.0074, P_{\max,4} = 0.3381, P_{\max,5} = 0.3744$ *и* $p = \max(0.3381; 0.3744) = 0.3744$.

После шага 4: $p_u = 0.6531$.

Шаг 5: оценка минимальной энтропии равна:

$$H_{\min} = -log_2(0.6531) = 0.6146.$$

5.2.2.7 Тест множественного прогнозирования наиболее частого значения в окне (MultiMCW)

Общие сведения. `MultiMCW` -прогноз содержит несколько подпрогнозов, каждый из которых ставит целью угадать следующее выходное значение, основываясь на последних w выходных значениях. Каждый подпрогноз предсказывает значение, которое происходит чаще всего в этом окне из w предыдущих выходных значений. MultiMCW-прогноз содержит таблицу, хранящую количество раз, которое каждый из подпрогнозов был правильным, и использует подпрогноз с наиболее правильными предсказаниями для предсказывания следующего значения. MultiMCW-прогноз был разработан для случаев, когда наиболее частое значение меняется с течением времени, но по-прежнему остается относительно постоянным при разумных длинах последовательностей.

Входные данные: $S = (s_1, \ldots, s_n)$; $\quad s_i \in A = \{x_1, \ldots, x_k\}$ $(k \geqslant 2)$.

bf Алгоритм теста множественного прогнозирования наиболее частого значения в окне (тест `MultiMCW`) состоит из следующих шагов.

1) Определить размеры окон: $w_1 = 63, w_2 = 255, w_3 = 1023, w_4 = 4095$ и $N = n - w_1$. Создать массив `correct` из N булевых значений, каждое из которых равно 0.

2) Пусть `scoreboard` — массив из четырех счетчиков, каждый из которых равен 0. Пусть `frequent` – массив из четырех значений, каждое из которых равно Null. Пусть `winner`=1.

3) Для i от $w_1 + 1$ до n:

а) Для j от 1 до 4:

(a) Если $i > w_j$, пусть $frequent[j]$ — наиболее частое значение в $(S_{i-w_j}, S_{i-w_j+1}, \ldots, S_{i-1})$. В случае равенства, $frequent[j]$ присваивается наиболее частое значение, которое появилось последним.

(b) Иначе, положить $frequent[j] = Null$.

б) Пусть $prediction = frequent_{winner}$.

в) Если $(prediction = s_i)$, то положить $correcti - w_1 = 1$.

г) Обновить $scoreboard$. Для j от 1 до 4:

(a) Если $(frequentj = s_i)$:

A. то $scoreboardj = scoreboardj + 1$.

B. Если $scoreboardj \geqslant scoreboard[winner]$, то $winner = j$.

4) Пусть C — количество единиц в `correct`.

5) Вычислить общую эффективность прогноза: $P_{global} = \frac{C}{N}$.

Верхняя граница с 99.0% ой доверительной вероятностью:

$$P_{global2} = \{1 - 0.01^{1/N}, \text{если} P_{global} = 0; \quad \min(1, P_{global} + 2.576\sqrt{\frac{P_{global}(1 - P_{global})}{n-1}}), \text{иначе}\},$$

где 2.576 соответствует квантилю нормального распределения.

6) Вычислить локальную эффективность прогноза на основании наиболее длинной продолжительности правильных предсказаний. Пусть r на единицу больше, чем длина наиболее длинной продолжительности единиц в `correct`. С использованием бинарного поиска решить относительно P_{local}:

$$0.99 = \frac{1 - P_{local}x}{(r + 1 - rx)q} \times \frac{1}{x^{N+1}},$$

где $q = 1 - P_{local}$ и $x = x_{10}$, получается итерационным способом из

$$x_j = 1 + qP_{local}^r x_{j-1}^{r+1},$$

для j от 1 до 10, $x_0 = 1$.

7) Оценка минимальной энтропии является отрицательным логарифмом большего из показателей:

$$H_{\min} = -log_2(\max(P_{global2}, P_{local}, \frac{1}{k})).$$

Условия применения: рекомендуется $n \geqslant 1000000$.

5.2.2.8 Тест прогнозирования задержки

Общие сведения. Прогноз задержки содержит несколько подпрогнозов, каждый из которых предсказывает следующий выход, основываясь на заданной задержке. Прогноз задержки содержит таблицу, хранящую количество раз, которое каждый из подпрогнозов был правильным, и использует подпрогноз с наиболее правильными предсказаниями для предсказывания следующего значения.

Входные данные: $S = (s_1, \ldots, s_n); \quad s_i \in A = x_1, \ldots, x_k$.

bf Алгоритм теста прогнозирования задержки (тест `LPE`) состоит из следующих шагов.

1) Пусть $D = 128$ и $N = n-1$. Пусть `lag` — массив из D значений, каждое из которых равно `Null`. Создать массив *correct* из $N0$ булевых значений, каждое из которых равно 0.

2) Пусть `scoreboard` — массив из D счетчиков, каждый из которых равен 0. Пусть $winner = 1$.

3) Для i от 2 до n:

а) Для d от 1 до D:

(a) Если $d < i$, то $lag_d = s_{i-d}$,

(b) Иначе, $lag_d = Null$.

б) Пусть $prediction = lag_{winner}$.

в) Если $(prediction = s_i)$, то $correct_{i-1} = 1$.

г) Обновляеv `scoreboard`. Для d от 1 до D:

(a) Если $lag_d = s_i$, то:

A. $scoreboard_d = scoreboard_d + 1$.

B. Если $scoreboard_d \geqslant scoreboard_{winner}$, то $winner = d$.

4) Пусть C — количество единиц в `correct`.

5) Вычислить общая эффективность прогноза: $P_{global} = \frac{C}{N}$. Верхняя граница с доверительной вероятностью 0,99:

$$P_{global2} = \{1 - 0.01^{1/N},\ \text{если} P_{global} = 0; \quad \min(1, P_{global} + 2.576\sqrt{\frac{P_{global}(1-P_{global})}{N-1}}, \text{иначе}\},$$

где 2.576 соответствует квантилю нормального распределения $u_{(1-0.005)}$.

6) Вычислить локальную эффективность прогноза, основанную на наиболее длинной продолжительности правильных предсказаний. Пусть r на единицу больше, чем длина наиболее длинной продолжительности единиц в `correct`. С использованием бинарного поиска решить относительно P_{local}:

$$0.99 = \frac{1 - P_{local}x}{(r+1-rx)q} \times \frac{1}{x^{N+1}},$$

где $q = 1 - P_{local}$ и $x = x_{10}$, получается итерационным способом из

$$x_j = 1 + qP_{local}^r x_{j-1}^{r+1}$$

для j от 1 до 10, и $x_0 = 1$.

7) Оценка минимальной энтропии является отрицательным логарифмом большего из показателей:

$$H_{\min} = -log_2(\max(P_{global2}, P_{local}, \frac{1}{k})).$$

Условия применения: рекомендуется $n \geqslant 1000000$.

5.2.2.9 Тест множественного прогнозирования моделей Маркова с подсчетом

Общие сведения. `MultiMMC` -прогноз состоит из множества подпрогнозов моделей Маркова с подсчетом. Каждый MMC-прогноз записывает наблюдаемые частоты для переходов от одного выходного значения к последующему выходному значению (в отличие от вероятности перехода, как в типичной модели Маркова) и делает прогноз, основываясь на наиболее частом наблюдаемом переходе из текущего выходного значения. MultiMMC содержит D MMC-подпрогнозов, работающих параллельно, по одному для каждой глубины от 1 до D. Например, `MMC` с глубиной 1 создает модель первого порядка, в то время как `MMC` с глубиной D создает модель порядка D. `MultiMMC` содержит таблицу, хранящую количество раз, которое каждый из `MMC`-подпрогнозов был правильным, и использует подпрогноз с наиболее правильными предсказаниями для предсказывания следующего значения.

Входные данные: $S = (s_1, \ldots, s_L); s_i \in A = x_1, \ldots, x_k$.

Алгоритм теста множественного прогнозирования моделей Маркова (тест `MultiMMC`) состоит из следующих шагов.

1) Пусть $D = 16$ и $N = n - 2$. Пусть `subpredict` — массив из D значений, каждое из которых равно $Null$. Создать массив correct из N булевых значений, каждое из которых равно 0. Пусть записи представляют собой массив значений D, каждое из которых инициализировано 0, и пусть `maxEntries` = 100000.

2) Для d от 1 до D установить M_d — массив счетчиков, где $M_d[x, y]$ обозначает число наблюдаемых переходов от выходного значения x к выходному значению y для MMC порядка d.

3) Пусть `scoreboard` — массив из D счетчиков, каждый из которых равен 0. Полагаем $winner = 1$.

4) Для i от 3 до L:

a) Для d от 1 до D:

(a) Если $d < i - 1$, то:

A. Если $[(s_{i-d-1}, \ldots, s_{i-2}), s_{i-1}] \in M_d$, то увеличить $MMC_d[(S_{i-d-1}, \ldots, S_{i-2}), S_{i-1}]$ на 1.

B. Иначе, если $entries_d < maxEntries$, то счетчик для $[(s_{i-d-1}, \ldots, s_{i-2}), s_{i-1}]$ установлен: $M_d[(s_{i-d-1}, \ldots, s_{i-2}), s_{i-1}] = 1$ и увеличить $entries_d$ на 1.

5) Для d от 1 до D:

a) Если $d < i$, то находим значение y, которое соответствует наибольшему значению $M_d[(S_{i-d}, \ldots, S_{i-1}), y]$ и обозначается как $y_{\max}$. Пусть $subpredict_d = y_{\max}$. Если все возможные значения $M_d[(S_{i-d}, \ldots, S_{i-1}), y]$ равны 0, то положить $subpredict_d = Null$.

6) Пусть $prediction = subpredict_{winner}$.

7) Если $prediction = s_i$, то $correct_{i-2} = 1$.

8) Обновить $scoreboard$. Для d от 1 до D:

a) Если $(subpredict_d = s_i)$:

(a) $scoreboard_d = scoreboard_d + 1$.

(b) Если $scoreboard_d \geqslant scoreboard_{winner}$, то $winner = d$.

9) Пусть C — количество единиц в `correct`.

10) Вычислить общую эффективность прогноза: $P_{global} = \dfrac{C}{N}$. С доверительной вероятностью 0.99 верхняя граница равна:

$$P_{global2} = \{1 - 0.01^{1/N}, \text{если} P_{global} = 0; \quad \min(1, P_{global} + 2.576\sqrt{\frac{P_{global}(1 - P_{global})}{N - 1}}), \text{иначе}\},$$

где 2.576 соответствует квантилю нормального распределения $u_{0.995}$.

11) Вычислить локальную эффективность прогноза, основанную на наиболее длинной продолжительности правильных предсказаний. Пусть r на единицу больше, чем длина наиболее длинной продолжительности единиц в `correct`. С использованием бинарного поиска решить относительно P_{local}:

$$0.99 = \frac{1 - P_{local}x}{(r + 1 - rx)q} \times \frac{1}{x^{N+1}},$$

где $q = 1 - P_{local}$ и $x = x_{10}$, получается итерационным способом из

$$x_j = 1 + qP_{local}^r x_{j-1}^{r+1}$$

для j от 1 до 10, и $x_0 = 1$.

12) Оценка минимальной энтропии является отрицательным логарифмом большего из показателей:

$$H_{\min} = -log_2(\max(P_{global2}, P_{local}, \frac{1}{k})).$$

Условия применения: рекомендуется $n \geqslant 1000000$.

5.2.2.10 Тест прогнозирования LZ78Y

Общие сведения. Прогноз содержит словарь строк, которые были добавлены в словарь к текущему моменту, и продолжает добавление новых строк в словарь до тех пор, пока

словарь не достигнет своей максимальной емкости. Каждый раз, когда обрабатывается выборка, каждая подстрока из последних B выборок обновляет словарь или добавляется в словарь.

Входные данные: $S = (s_1, \ldots, s_n)$; $\quad s_i \in A = x_1, \ldots, x_k$.

Алгоритм теста прогнозирования LZ78Y (тест `LZ78Y`) состоит из следующих шагов.

1) Пусть $B = 16$ и $N = n - B - 1$. Создать массив `correct` из N булевых значений, каждое из которых равно 0. Пусть $maxDictionarySize = 65536$.

2) Пусть D — пустой словарь. Пусть $dictionarySize = 0$.

3) Для i от $B + 2$ до n:

а) Для j от B до 1:

(a) Если $(s_{i-j-1}, \ldots, s_{i-2})$ не содержится в D и $dictionarySize < maxDictionarySize$, то:

A. $D[s_{i-j-1}, \ldots, s_{i-2}]$ добавить в словарь.

B. $D[s_{i-j-1}, \ldots, s_{i-2}][s_{i-1}] = 0$.

C. $dictionarySize = dictionarySize + 1$.

(b) Если $(s_{i-j-1}, \ldots, s_{i-2})$ содержится в D, то:

A. $D[s_{i-j-1}, \ldots, s_{i-2}][s_{i-1}] = D[s_{i-j-1}, \ldots, s_(i-2)][s_{i-1}] + 1$.

б) Использовать словарь для предсказания следующего значения s_i. Пусть $prediction = Null$, и пусть $maxcount = 0$. Для j от B до 1:

(a) Пусть $prev = (s_{i-j}, \ldots, s_{i-1})$.

(b) Если $prev$ содержится в словаре, то найти $y \in x_1, \ldots, x_k$, который имеет наибольшее значение $D[prev][y]$. In the event of a tie, let the y be the symbol with the higher byte value. For example, if D[prev][1] and D[prev][5] both have the highest value, then y = 5.

(c) Если $D[prev][y] > maxcount$:

A. $prediction = y$.

B. $maxcount = D[prev][y]$.

в) Если $(prediction = s_i)$, то $correct_{i-B-1} = 1$.

4) Вычислить эффективность прогноза: $P_{global} = \dfrac{C}{N}$. С доверительной вероятностью 0.99 верхняя граница равна:

$$P_{global2} = \{1 - 0.01^{1/N}, \text{ если} P_{global} = 0; \quad \min(1, P_{global} + 2.576\sqrt{\frac{P_{global}(1 - P_{global})}{N - 1}}), \text{ иначе,}$$

где 2.576 соответствует квантилю нормального распределения $u_{0.995}$.

5) Вычислить локальную эффективность прогноза, основанную на наиболее длинной продолжительности правильных предсказаний. Пусть r на единицу больше, чем длина наиболее длинной продолжительности единиц в `correct`. С использованием бинарного поиска решить относительно P_{local}:

$$0.99 = \frac{1 - P_{local}x}{(r + 1 - rx)q} \times \frac{1}{x^{N+1}},$$

где $q = 1 - P_{local}$ и $x = x_{10}$, получается итерационным способом из

$$x_j = 1 + qP_{local}^r x_{j-1}^{r+1}$$

для j от 1 до 10, и $x_0 = 1$.

6) Оценка минимальной энтропии является отрицательным логарифмом большего из показателей:

$$H_{\min} = -log_2(\max(P_{global2}, P_{local}, \frac{1}{k})).$$

Условия применения: рекомендуется $n \geqslant 1000000$.

Приложение А
Форма протокола

Экз. № {Поле 1}

ПРОТОКОЛ № {Поле 2} **от** {Поле 3}

Источник случайности: {Поле 4}

Выходные последовательности:

Название	Режим работы	Число наблюдений	Размер наблюдения в битах

1. Проверка независимости и одинаковой распределённости

Тесты	Последовательности		
	{Назв. 1}	...	{Назв. N}
Тесты перестановок	{Поле 5}	...	{Поле 5}
Хи-квадрат тест для проверки независимости	{Поле 5}	...	{Поле 5}
Хи-квадрат тест для проверки одинаковой распределённости	{Поле 5}	...	{Поле 5}

Выходная последовательность источника случайности [в режимах ...] {прошла/не прошла} проверку на независимость и одинаковую распределённость.

2. Оценивание энтропии источника случайности

Тесты	Последовательности		
	{Назв. 1}	...	{Назв. N}
Тест н.о.р. случныйных величин	{Поле 6}	...	{Поле 6}
Частотный тест	{Поле 6}	...	{Поле 6}
Тест коллизий	{Поле 6}	...	{Поле 6}
Тест марковской зависимости	{Поле 6}	...	{Поле 6}
Тест сжатия	{Поле 6}	...	{Поле 6}
Тест частичных коллекций	{Поле 6}	...	{Поле 6}
Тест наибольшей повторяющейся подстроки	{Поле 6}	...	{Поле 6}
Тест MultiMCW	{Поле 6}	...	{Поле 6}
Тест прогнозирования задержки	{Поле 6}	...	{Поле 6}
Тест MultiMMC	{Поле 6}	...	{Поле 6}
Тест прогнозирования LZ78Y	{Поле 6}	...	{Поле 6}

Оценки минимальной энтропии в битах на наблюдение:

Режим работы	Оценка мин. энтропии

Эксперт(ы),
{Поле 7} {Поле 8}

В поле 1 указывается номер экземпляра протокола.

В поле 2 указывается номер, однозначно идентифицирующий протокол.

В поле 3 указывается дата составления протокола.

В поле 4 указывается название источника случайности, как заявлено производителем. Далее в таблице описываются исследуемые выходные последовательности источника случайности. При отсутствии различных режимов работы соответствующий столбец исключается.

В полях 5 указываются результаты проверки соответствующими статистическими тестами: «успешно» — тест успешно пройден, «не успешно» — тест не пройден, «недостаточно данных» — тест не выполнен по причине недостаточности данных.

Далее описываются выводы по результатам проверки независимости и одинаковой распределённости, прошла или не прошла выходная последовательность источника случайности данную проверку. При наличии различных режимов работы источника случайности, выводы описываются по каждому режиму работы.

В полях 6 указываются оценки минимальной энтропии, полученные в соответствующем тесте. Если какой-либо из тестов не выполнен по причине недостаточности данных, то в поле 6 указывается «недостаточно данных».

Далее для каждого режима работы указывается итоговая оценка минимальной энтропии. Если источник случайности работает только в одном режиме, то указывается только одно число.

В полях 7, 8 указываются соответственно должность и Ф.И.О. экспертов.

Библиография

[1] Barker E. Recommendation for the Entropy Sources Used for Random Bit Generation / M. Turan, E. Barker, J. Kelsey, K. McKay, M. Baish, M. Boyle// This publication is available free of charge from: https://doi.org/10.6028/NIST.SP.800-90B. — 2018. — 84 p.

[2] Кнут Д. Искусство программирования / Д. Кнут. — 3-е изд. Москва: Вильямс, 2011. — 832 с.

[3] BZ2 compression algorithm [Электронный ресурс]. — Режим доступа: http://www.bzip.org/. Дата доступа: 17.12.2015.